

Historic, Archive Document

Do not assume content reflects current scientific knowledge, policies, or practices.



United States
Department of
Agriculture

Forest Service

Rocky Mountain
Forest and Range
Experiment Station

Fort Collins,
Colorado 80526

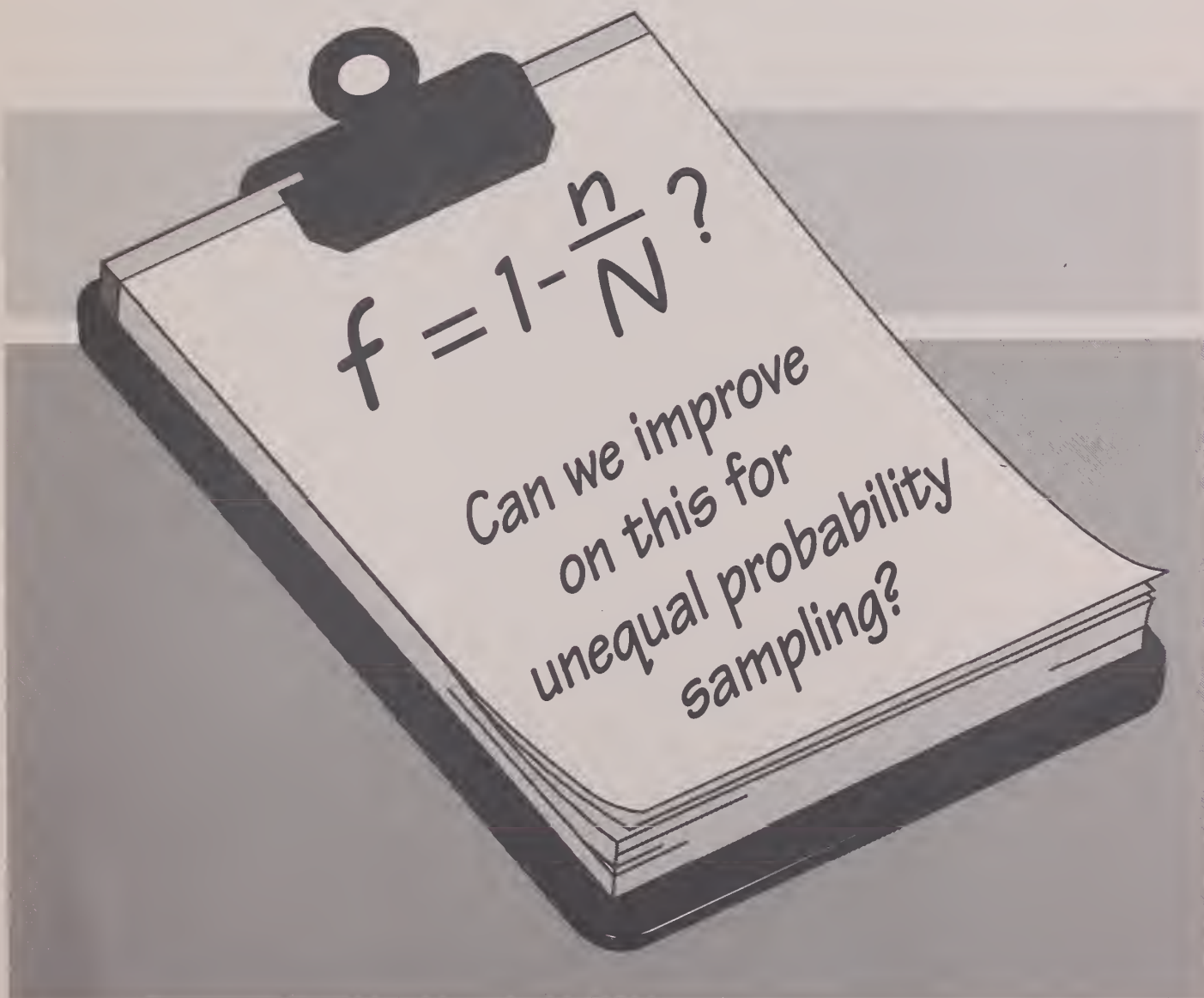
Research Paper
RM-RP-329



Finite Population Corrections of the Horvitz-Thompson Estimator and their Application in Estimating the Variance of Regression Estimators

Z. Ouyang
H.T. Schreuder
D.C. Boes

RM-RP-329
A 5-08



Ouyang, Z., Schreuder, H. T., and Boes, D. C. 1997. Finite Population Corrections of the Horvitz-Thompson Estimator and Their Application in Estimating the Variance of Regression Estimators. Res. Pap. RM-RP-329. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station. 8 p.

ABSTRACT

Kott (1988) first derived finite population corrections (fpcs) for the Horvitz-Thompson estimator under unequal probability sampling without replacement. No numerical comparison of variances with the fpcs was made. We derive some of the fpcs again in a new and more intuitive way without resorting to large sampling theory. We demonstrate how to incorporate these fpcs with the jackknife variance estimator of the generalized regression estimator (Särndal 1980). Simulation results showed that adjustments by these fpcs resulted in variances with considerable less bias. One of the fpcs is also robust.

Key words. unequal probability sampling, pps sampling, bias correction, variance estimation of regression estimator

You may order additional copies of this publication by sending your mailing information in label form through one of the following media. Please send the publication title and report number.

Telephone (970) 498-1719

DG message R.Schneider:S28A

FAX (970) 498-1660

E-mail /s=r.schneider/ou1=s28a@mhs-fswa.attmail.com

Mailing Address Publications Distribution
Rocky Mountain Forest and Range
Experiment Station
3825 E. Mulberry Street
Fort Collins, CO 80524

Finite Population Corrections of the Horvitz-Thompson Estimator and their Application in Estimating the Variance of Regression Estimators

Z. Ouyang¹

H.T. Schreuder²

D.C. Boes³

¹Formerly Post Doctoral Fellow, Department of Statistics, Colorado State University, Fort Collins, CO. Now Biostatistician, Otsuka American Pharmaceutical, 2440 Research Blvd., Rockville, MD 20850

²Mathematical Statistician, Multiresource Inventory Techniques, Rocky Mountain Station, 240 West Prospect, Fort Collins, CO 80526.

³Professor, Department of Statistics, Colorado State University, Fort Collins, CO.

Contents

	Page
Introduction	1
Finite Population Correction	1
Numerical Comparison of the fpc	3
Finite Population Correction of a Jackknife Variance	5
Estimator for a Regression Estimator	5
Summary	7
References	8

Finite Population Corrections of the Horvitz-Thompson Estimator and their Application in Estimating the Variance of Regression Estimators

Z. Ouyang
H. T. Schreuder
D. C. Boes

Introduction

The use of the Horvitz-Thompson (HT) estimator (\hat{Y}_{HT}) with probability proportional to size (pps) sampling without replacement (ppswor) is well established (Brewer and Hanif 1983). Brewer and Hanif (1983) list 50 different ppswor procedures but noted that with most of the procedures the second order inclusion probabilities are difficult to calculate. Consequently, the elegant classical variance estimators of the Horvitz-Thompson estimator cannot be used. Sometimes, a variance estimator for \hat{Y}_{pps} is used assuming sampling with probability proportional to size with replacement. But, this variance estimator, denoted by $v(\hat{Y}_{pps})$, is biased upward (Wolter 1985, Schreuder and Ouyang, 1992) as expected.

For a finite population with variable of interest y and the following approximate model for y :

$$y_i = bx_i + e_i\sigma_i, \quad i = 1, \dots, N, \quad (1.1)$$

where the x_i are known positive numbers satisfying the constraint

$$\max[x_1, \dots, x_N] < (1/n) \sum_{i=1}^N x_i,$$

and n is the sample size, the e_i are uncorrelated random variables with mean 0 and variance 1 and each σ_i is a positive function of x_i which may or may not be specified. To improve on $v(\hat{Y}_{pps})$ as a variance estimator of \hat{Y}_{HT} by utilizing the information provided by the model, consideration of a finite population correction (fpc) is necessary.

$$Pi = x_i / \sum_{j=1}^N x_j$$

The probability of selecting unit i in each draw and $\pi_i = n\pi_i$ is the inclusion probability of unit i .

Finite Population Correction

Fpcs for the Horvitz-Thompson estimator under ppswor was first addressed by Wolter (1985) who suggested using

$$1 - \frac{1}{n} \sum_{i \in s} \Pi_i$$

as an fpc without providing justification. Kott (1988) derived fpc of the Horvitz-Thompson estimator under ppswor by replacing the design variance of \hat{Y}_{HT} with its model variance and the design expectation of $v(\hat{Y}_{pps})$ with its model expectation and suggested that their ratio could be used as an fpc. Although Kott's original fpc is complex, he derived a simple expression for the fpc by taking its limit. However, replacing design variance and design expectation by model variance and model expectation is not easily justified.

The fpc in simple random sampling without replacement (srswor) is:

$$1 - \frac{n}{N} = [(N-n)/(Nn)] [1/(N-1)] \sum_{i=1}^N (y_i - \bar{Y})^2 / \left\{ 1/[n(N-1)] \sum_{i=1}^N (y_i - \bar{Y})^2 \right\}, \quad (2.1)$$

where the numerator is equal to the variance of \bar{y}_s (sample mean) under srswor and the denominator is equal to the expectation of

$$\frac{n}{n(n-1)} \sum_{i \in s} (y_i - \bar{y}_s)^2$$

under srswor.

$$\frac{n}{n(n-1)} \sum_{i \in s} (y_i - \bar{y}_s)^2$$

is the variance of \bar{y}_s under simple random sampling with replacement (srswr). The fpc in srswor can be expressed as:

$$1 - \frac{n}{N} = E(\bar{y}_s - \bar{Y})^2 / E[v_w(\bar{y}_s)] \quad (2.2)$$

where $v_w(\bar{y}_s)$ is the variance estimator of \bar{y}_s under srswr. Using the fpc in srswor by (2.2) suggests the use of

$$\left(1 - \frac{n}{N}\right) v_w(\bar{y}_s)$$

as a variance estimator of \bar{y}_s in srswor. Generalizing this idea to sampling ppswor, an fpc (denoted by $1 - f_{pps}$) can be written as:

$$1 - f_{pps} = E(\hat{Y}_{HT} - Y)^2 / E(v_w(\hat{Y}_{pps})) \quad (2.3)$$

where

$$v_w(\hat{Y}_{pps}) = \frac{1}{n(n-1)} \sum_{i \in s} (y_i / p_i - \hat{Y}_{pps})^2 \quad (2.4)$$

is the variance estimator of \hat{Y}_{pps} under ppswr and

$$\hat{Y}_{pps} = \frac{1}{n} \sum_{i \in s} y_i / p_i. \quad (2.5)$$

$(1 - f_{pps}) v_w(\hat{Y}_{pps})$ can be used as a variance estimator of \hat{Y}_{HT} if f_{pps} is known. This variance estimator will correct the upward bias of $v_w(\hat{Y}_{pps})$ as a variance estimator of \hat{Y}_{HT} . Based on (2.3), $1 - f_{pps}$ is derived. The design expectation of $v_w(\hat{Y}_{pps})$ given in (2.4) is:

$$Ev_w(\hat{Y}_{pps}) = \frac{1}{n(n-1)} E \left[\sum_{i \in s} (y_i / p_i - \hat{Y}_{pps})^2 \right] \quad (2.6)$$

$$\begin{aligned} &= \frac{1}{n(n-1)} E \left[\sum_{i \in s} y_i^2 / p_i^2 - 2\hat{Y}_{pps} \sum_{i \in s} y_i / p_i + n\hat{Y}_{pps}^2 \right] \\ &= \sum_{i=1}^N y_i^2 / \Pi_i - \frac{1}{n-1} \sum_{i \neq j} y_i y_j \Pi_{ij} / (\Pi_i \Pi_j) \end{aligned}$$

where Π_{ij} is the joint probability of selecting units i and j .

The variance of the HT estimator is:

$$Var(\hat{Y}_{HT}) = \sum_{i=1}^N \frac{1 - \Pi_i}{\Pi_i} y_i^2 + \sum_{i \neq j}^N \frac{\Pi_{ij} - \Pi_i \Pi_j}{\Pi_i \Pi_j} y_i y_j \quad (2.7)$$

Substituting (2.6) and (2.7) into (2.3) yields

$$\begin{aligned} 1 - f_{pps} &= \left[\sum_{i=1}^N \frac{1 - \Pi_i}{\Pi_i} y_i^2 + \sum_{i \neq j}^N \frac{\Pi_{ij} - \Pi_i \Pi_j}{\Pi_i \Pi_j} y_i y_j \right] / \\ &\quad \left[\sum_{i=1}^N y_i^2 / \Pi_i - \frac{1}{n-1} \sum_{i \neq j}^N y_i y_j \Pi_{ij} / \Pi_i \Pi_j \right]. \end{aligned} \quad (2.8)$$

The fpc expression given in (2.8) cannot be used directly since the second-order inclusion probabilities on the right-hand side are unknown. But if the second-order inclusion probabilities were known, $Var(\hat{Y}_{HT})$ could be estimated directly without resorting to the use of fpc. By using the model in (1.1), the second order inclusion probabilities in (2.8) are eliminated. If E_ξ is the expectation operator under model (1.1), then:

$$\begin{aligned} 1 - f_{pps} &\doteq \frac{E_\xi \left[\sum_{i=1}^N \frac{(1 - \Pi_i)}{\Pi_i} y_i^2 + \sum_{i \neq j}^N \frac{\Pi_{ij} - \Pi_i \Pi_j}{\Pi_i \Pi_j} y_i y_j \right]}{E_\xi \left[\sum_{i=1}^N y_i^2 / \Pi_i - \frac{1}{n-1} \sum_{i \neq j}^N y_i y_j \Pi_{ij} / \Pi_i \Pi_j \right]} \\ &= \frac{\sum_{i=1}^N \frac{1 - \Pi_i}{\Pi_i} (b^2 x_i^2 + \sigma_i^2) + \sum_{i \neq j}^N \frac{\Pi_{ij} - \Pi_i \Pi_j}{\Pi_i \Pi_j} b^2 x_i x_j}{\sum_{i=1}^N (b^2 x_i^2 + \sigma_i^2) / \Pi_i - \frac{1}{n-1} \sum_{i \neq j}^N b^2 x_i x_j \Pi_{ij} / \Pi_i \Pi_j} \end{aligned} \quad (2.9)$$

$$\begin{aligned}
&= \frac{\sum_{i=1}^N \frac{1-\Pi_i}{\Pi_i} (\Pi_i^2 + c\sigma_i^2) + \sum_{i \neq j}^N \frac{\Pi_{ij} - \Pi_i \Pi_j}{\Pi_i \Pi_j} \Pi_i \Pi_j}{\sum_{i=1}^N (\Pi_i^2 + c\sigma_i^2) / \Pi_i - \frac{1}{n-1} \sum_{i \neq j}^N \Pi_i \Pi_j \Pi_{ij} / \Pi_i \Pi_j} \\
&= \frac{c \sum_{i=1}^N \frac{1-\Pi_i}{\Pi_i} \sigma_i^2 + \sum_{i=1}^N (\Pi_i - \Pi_i^2) + \sum_{i \neq j}^N (\Pi_{ij} - \Pi_i \Pi_j)}{c \sum_{i=1}^N \sigma_i^2 / \Pi_i + \sum_{i=1}^N \Pi_i - \frac{1}{n-1} \sum_{i \neq j}^N \Pi_{ij}},
\end{aligned}$$

(2.9 continued)

where

$$c = n^2 / \left(b \sum_{i=1}^N x_i \right)^2.$$

Because:

$$\sum_{i \neq j}^N \Pi_{ij} = \sum_{i=1}^N \sum_{j(\neq i)}^N \Pi_{ij} = (n-1) \sum_{i=1}^N \Pi_i = (n-1)n, \quad (2.10)$$

and

$$\begin{aligned}
\sum_{i \neq j}^N \Pi_i \Pi_j &= \sum_{i=1}^N \sum_{j(\neq i)}^N \Pi_i \Pi_j = \sum_{i=1}^N \Pi_i \sum_{j(\neq i)}^N \Pi_j \\
&= \sum_{i=1}^N \Pi_i (n - \Pi_i) = n^2 - \sum_{i=1}^N \Pi_i^2,
\end{aligned} \quad (2.11)$$

substituting (2.10) and (2.11) into (2.9) leads to:

$$\begin{aligned}
1 - f_{pps} &\doteq \frac{c \sum_{i=1}^N \frac{1-\Pi_i}{\Pi_i} \sigma_i^2 + \sum_{i=1}^N (\Pi_i - \Pi_i^2) + \left[n(n-1) - n^2 + \sum_{i=1}^N \Pi_i^2 \right]}{c \sum_{i=1}^N \sigma_i^2 / \Pi_i + \sum_{i=1}^N \Pi_i - \frac{1}{n-1} n(n-1)} \\
&= \frac{\sum_{i=1}^N \frac{1-\Pi_i}{\Pi_i} \sigma_i^2}{\sum_{i=1}^N \frac{\sigma_i^2}{\Pi_i}} \\
&= 1 - \frac{\sum_{i=1}^N \sigma_i^2}{\sum_{i=1}^N \Pi_i}.
\end{aligned} \quad (2.12)$$

Kott (1988) derived (2.12) for large n . But (2.12) is also good for small n and the above derivation is more straightforward. If Π_i^2 is proportional to σ_i^2 , then:

$$1 - f_1 = 1 - \frac{1}{n} \sum_{i=1}^N \Pi_i^2. \quad (2.13)$$

Note that

$$\sum_{i=1}^N \Pi_i^2$$

can be "estimated" by

$$\sum_{i \in S} \Pi_i^2 / \Pi_i = \sum_{i \in S} \Pi_i,$$

so:

$$1 - f_2 = 1 - \frac{1}{n} \sum_{i \in S} \Pi_i. \quad (2.14)$$

Wolter (1985) also suggested using $1-f_2$ as an fpc, even though he did not provide any rationale.

Using the harmonic mean to replace the arithmetic mean in (2.14), a somewhat conservative fpc is:

$$1 - f_3 = 1 - n / \sum_{i \in S} \frac{1}{\Pi_i}. \quad (2.15)$$

Kott (1988) also derived f_3 under the assumption $\sigma_i^2 \propto \Pi_i$ but based on our derivation, if Π_i is proportional to σ_i^2 , then (2.12) gives:

$$1 - f_4 = 1 - n / N \quad (2.16)$$

as a finite population correction directly: $1-f_4$ can also be estimated by $1-f_3$. Equation (2.16) is an upper bound of $1-f_1$ because:

$$n^2 = \left(\sum_{i=1}^N \Pi_i \right) \left(\sum_{i=1}^N \Pi_i \right) \leq N \sum_{i=1}^N \Pi_i^2.$$

So (2.16) is also a conservative fpc compared to $1-f_1$.

Numerical Comparison of the fpc

We used 3 data sets from forest surveys to compare the fpc in a simulation study. The first data set,

LOBK2, consisted of 1795 trees with variables total bole volume in cubic feet (v_c) and tree diameter at breast height squared in inches times total height in feet (d^2h). For each tree the value v_c was computed from the equation:

$$v_c = -0.734 + 0.002129 d^2h + e_i$$

with error e_1 approximately distributed as $N(0, \sigma^2(d^2h)^2)$, where $\hat{\sigma}^2 = 0.00001$ (Schreuder and Ouyang 1992). A very strong linear relationship existed between v_c and d^2h . The second data set, NY, consisted of plot volume on 622 remeasured plots in New York. The variables were gross cubic foot volume per acre of live trees (GAL) at time t and $t-1$. Simple linear regression of GAL_t on GAL_{t-1} was:

$$GAL_t = 1053.706 + 0.5313058 GAL_{t-1} + e_2$$

with $R^2 = 0.2156$. The residuals e_2 might be roughly proportional to GAL_{t-1} (Schreuder et al. 1990). The third population was RG5 with loblolly pine radial growth data at 2 points in time (Zahner et al. 1989). A weak linear relationship existed between the variables and the variance structure of the dependent variable y given x was unclear.

The entire NY population and random samples of size 500 from population LOBK2 and RG5 were used as populations, since population sizes should be small for the fpc to matter. Random samples totaling 1000 of size 30 were drawn from each population. The mean of the square root of each variance estimator obtained in the simulation was then expressed as percent of simulation standard error. Three unequal probability sampling strategies were used. The first one was a stratified sampling from the cumulated x 's (spscx): Sort the population by x -value. Then divide the sum of the x 's,

$$XT = \sum_{i=1}^n x_i,$$

by the derived sample size n , and from n strata $[(XT/n) i, (XT/n) (i+1)]$ $i=0, 1, \dots, (n-1)$, select one sample unit at random from each stratum. The second one was very similar to spscx, except $x_i^{1.5}$ was used instead of x_i in forming the strata. We call it spscxk. With the third method, called spsps sampling, one unit was selected from each stratum with probability proportional to its x , spscx and spsps used the same strata. Only the third one, spsps, was a true pps procedure. The other 2 were included to determine how sensitive the fpcs were to the assumption of y_i being proportional to π_i . To evaluate the fpcs:

$$v_w = \frac{1}{n(n-1)} \sum_{i \in S} (y_i / p_i - \hat{Y}_{pp2})^2, \quad v_1 = (1 - f_1) v_w, \\ v_2 = (1 - f_2) v_w, \quad v_3 = (1 - f_3) v_w, \quad v_4 = (1 - f_4) v_w,$$

were used in the simulation study. The mean square roots of the variances expressed as percent of simulation standard errors (table 1) indicated that adjusting v_w by the fpc usually results in estimators that are less biased than v_w .

Based on the simulation, v_1 was the best estimator. This was expected for LOBK2, since v_1 was derived based on equation (1.1). But v_1 was also best for NY. The variance of the error term for NY was proportional to something between GAL_{t-1} to $(GAL_{t-1})^2$. v_1 and v_2 were best in RGS, which suggested that v_1 was very robust. v_3 was basically more conservative than v_1 in all cases. v_3 and v_4 behaved quite similarly since

$$\sum_{i \in S} 1/\pi_i$$

was an estimator of n/N but v_4 tended to be a bit more stable since $1-n/N$ was a constant. v_1 was simi-

Table 1. Mean square root of variance estimator expressed as percent of the simulation standard error.

	LOBK2			NY			RG5		
	spscx	spscxk	spps	spscx	spscxk	spps	spscx	spscxk	spps
v_w	108.02	110.01	117.96	135.18	116.54	105.43	106.29	102.06	110.77
v_1	98.83	100.65	107.93	130.24	112.28	101.58	101.16	97.14	105.42
v_2	98.74	103.19	112.39	130.23	112.83	102.46	101.17	97.56	106.22
v_3	103.89	105.80	113.45	131.88	113.70	102.97	102.23	98.61	106.52
v_4	103.89	105.80	113.45	131.88	113.70	102.00	102.23	97.61	106.53

lar to and should be more stable than v_2 and was a somewhat less biased estimator than v_2 . In general, v_1 was the least biased variance estimator and v_w the most biased.

Finite Population Correction of a Jackknife Variance

Estimator for a Regression Estimator

Suppose that the finite population satisfies the following model:

$$y_i = a + bx_i + e_i, i = 1, \dots, N, \quad (4.1)$$

where a and b are unknown parameters for all i and the e_i are unknown random variables, with

$$E_{\xi} e_i = 0, E_{\xi} e_i e_j = \begin{cases} 0 & \text{if } i \neq j \\ \delta_i^2 \sigma^2 & \text{if } i = j \end{cases}, \quad (4.2)$$

where E_{ξ} denotes the expectation operator with respect to the model, σ^2 is an unknown parameter, and δ_i is a known or unknown function of the known values x_i . For populations that satisfy a model like (4.1), regression estimators are very efficient in estimating the population total Y .

Let $s = \{1, \dots, n\}$ be the index set of the sample. Define:

$$x_s = \begin{bmatrix} 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, y_s = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \Pi_s = \begin{bmatrix} \Pi_1 & & 0 \\ & \ddots & \\ 0 & & \Pi_n \end{bmatrix} \quad (4.3)$$

and:

$$\begin{bmatrix} \hat{a}_s \\ \hat{b}_s \end{bmatrix} = [x_s' \Pi_s^{-1} x_s]^{-1} x_s' \Pi_s^{-1} y_s. \quad (4.4)$$

Then, a general regression estimator (Särndal 1980) is:

$$\hat{Y}_{grpi} = N\hat{a}_s + \sum_{i=1}^N x_i \hat{b}_s + \sum_{i \in s} (y_i - \hat{a}_s - x_i \hat{b}_s) / \Pi_i. \quad (4.5)$$

In \hat{Y}_{grpi} , gr is the general regression estimator and Π_i^{-1} denotes weight Π_i^{-1} to estimate a and b in (4.4). Let the "census fit" of a and b be:

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = [X_N' X_N]^{-1} X_N' y_N, \quad (4.6)$$

where:

$$X_N = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}, y_N = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}. \quad (4.7)$$

Under some mild conditions (see Ouyang et al. 1991):

$$\hat{a}_s \doteq \hat{a}, \hat{b}_s \doteq \hat{b}. \quad (4.8)$$

Hence:

$$E(\hat{Y}_{grpi} - Y)^2 \doteq \text{Var} \left[\sum_{i \in s} \frac{y_i - \hat{a} - x_i \hat{b}}{\Pi_i} \right]. \quad (4.9)$$

Consider the jackknife variance estimator of \hat{Y}_{grpi} . Define the pseudovalues used in jackknifing by

$$\hat{\theta}_{\alpha} = n\hat{Y}_{grpi} - (n-1)\hat{Y}_{(\alpha)}, \quad (4.10)$$

where:

$$\hat{Y}_{(\alpha)} = N\hat{a}_{(\alpha)} + \sum_{i=1}^N x_i \hat{b}_{(\alpha)} + \left\{ n / (n-1) \right\} \sum_{\substack{i \in s \\ i \neq \alpha}} (y_i - \hat{a}_{(\alpha)} - x_i \hat{b}_{(\alpha)}) / \Pi_i \quad (4.11)$$

and $\hat{a}_{(\alpha)}$ and $\hat{b}_{(\alpha)}$ are obtained from (4.4) without the α -th observation. Define the jackknife estimator of Y by:

$$\hat{\theta} = 1/n \sum_{\alpha \in s} \hat{\theta}_{\alpha} \quad (4.12)$$

and the jackknife variance estimator of $\hat{\theta}$ by:

$$v(\hat{\theta}) = \{1/n(n-1)\} \sum_{\alpha \in s} (\hat{\theta}_{\alpha} - \hat{\theta})^2. \quad (4.13)$$

Notice that under the condition $\hat{a}_{(\alpha)} \doteq \hat{a}_s, \hat{b}_{(\alpha)} \doteq \hat{b}_s$, and the condition in (4.8), we have that $\hat{\theta}_{\alpha} - \hat{\theta}$ in (4.13) is approximately equal to:

$$\begin{aligned}
\hat{\theta}_\alpha - \frac{1}{n} \sum_{a \in S} \hat{\theta}_\alpha &\doteq -n \sum_{\substack{i \in S \\ i \neq \alpha}} (y_i - \hat{a}_s - x_i \hat{b}_s) / \Pi_i \\
&+ \frac{n-1}{n} \cdot \frac{n}{n-1} \sum_{a \in S} \left[\sum_{i \in S} (y_i - \hat{a}_s - x_i \hat{b}_s) / \Pi_i \right] \\
&= (n-1) \sum_{i \in S} (y_i - \hat{a}_s - x_i \hat{b}_s) / \Pi_i \\
&- n \sum_{\substack{i \in S \\ i \neq \alpha}} (y_i - \hat{a}_s - x_i \hat{b}_s) / \Pi_i \\
&= n(y_\alpha - \hat{a} - x_\alpha \hat{b}_s) / \Pi_\alpha - \sum_{i \in S} (y_i - \hat{a}_s - x_i \hat{b}_s) / \Pi_i \\
&\doteq n(y_\alpha - \hat{a} - x_\alpha \hat{b}) / \Pi_\alpha - \sum_{i \in S} (y_i - \hat{a} - x_i \hat{b}) / \Pi_i.
\end{aligned} \tag{4.14}$$

Thus, the jackknife variance estimator has the same form as (2.4), with y_i in (2.4) replaced by $(y_i - \hat{a} - x_i \hat{b})$. Note that:

$$y_i - \hat{a} - x_i \hat{b} = \hat{e}_i \doteq \hat{e}_i. \tag{4.15}$$

As pointed out by Cochran (1977, p. 257), it is often assumed in (4.2) that:

$$\delta_i = x_i^g, 1 \leq g \leq 2 \tag{4.16}$$

or

$$\delta_i = c_1 x_i + c_2 x_i^2 \quad (0 \leq c_1, c_2 \leq 1). \tag{4.17}$$

For $\delta_i = x_i^2$, we have:

$$y_i - \hat{a}_s - x_i \hat{b}_s \doteq e_i = \xi_i x_i \quad i = 1, \dots, N, \tag{4.18}$$

where $\xi_i \sim N(0, \sigma^2)$ and ξ_1, \dots, ξ_N are independent. The jackknife variance estimator has the following form:

$$\begin{aligned}
&\frac{1}{n(n-1)} \sum_{a \in S} \left[\sum_{i \in S} \left(\frac{n(y_\alpha - \hat{a}_s - x_\alpha \hat{b}_s) / \Pi_\alpha - \sum_{i \in S} (y_i - \hat{a}_s - x_i \hat{b}_s) / \Pi_i}{n-1} \right)^2 \right] \\
&\doteq \frac{1}{n(n-1)} \sum_{a \in S} \left[e_\alpha / p_\alpha - \sum_{i \in S} e_i / \Pi_i \right]^2.
\end{aligned} \tag{4.19}$$

To estimate $E(\hat{Y}_{grpi} - Y)^2$, which has the following form based on (4.9):

$$E(\hat{Y}_{grpi} - Y)^2 \doteq Var \left[\sum_{i \in S} e_i / \Pi_i \right], \tag{4.20}$$

the ξ_i s in (4.18) are different so that the fpcs developed for \hat{Y}_{HT} can not be used directly to adjust the jackknife variance estimator for \hat{Y}_{grpi} . To find the proper adjustment, we used y_i in (2.8) replaced by $y_i - \hat{a}_s - x_i \hat{b}_s$. Let E_ξ be the expectation operator under model (4.1) and (4.2). Since:

$$\begin{aligned}
E_\xi \left[\sum_{i \neq j} \frac{\Pi_{ij} - \Pi_i \Pi_j}{\Pi_i \Pi_j} (y_i - \hat{a}_s - x_i \hat{b}_s) (y_j - \hat{a}_s - x_j \hat{b}_s) \right] \\
\doteq E_\xi \left[\sum_{i \neq j} \frac{\Pi_{ij} - \Pi_i \Pi_j}{\Pi_i \Pi_j} e_i e_j \right] = 0
\end{aligned} \tag{4.21}$$

and

$$\begin{aligned}
E_\xi \left[\frac{1}{n-1} \sum_{i \neq j} \frac{\Pi_{ij}}{\Pi_i \Pi_j} (y_i - \hat{a} - x_i \hat{b}) (y_j - \hat{a} - x_j \hat{b}) \right] \\
\doteq E_\xi \left(\frac{1}{n-1} \sum_{i \neq j} \frac{\Pi_{ij}}{\Pi_i \Pi_j} e_i e_j \right) = 0.
\end{aligned} \tag{4.22}$$

Thus, the finite population correction factor in this case is approximated by:

$$\begin{aligned}
&\left[\sum_{i=1}^N \frac{1 - \Pi_i}{\Pi_i} (y_i - \hat{a} - x_i \hat{b})^2 \right] / \left[\sum_{i=1}^N (y_i - \hat{a} - x_i \hat{b})^2 / \Pi_i \right] \\
&\doteq \left[\sum_{i=1}^N \frac{1 - \Pi_i}{\Pi_i} (y_i - \hat{a}_s - x_i \hat{b}_s)^2 \right] / \left[\sum_{i=1}^N (y_i - \hat{a}_s - x_i \hat{b}_s)^2 / \Pi_i \right] \\
&\doteq \left[\sum_{i=1}^N \frac{1 - \Pi_i}{\Pi_i} \delta_i^2 \sigma_i^2 \right] / \sum_{i=1}^N \delta_i^2 \sigma_i^2 / \Pi_i
\end{aligned}$$

Thus, the fpc can be used here, with σ_i^2 in (2.12) replaced by $\delta_i^2 \gamma^2$, and fpc $(1-f_1)$, $(1-f_2)$, $(1-f_3)$, and $(1-f_4)$ can also be used here. For numerical comparisons, the 3 data sets and the 3 sampling schemes described in section 3 are used again. Besides the jackknife variance estimator (denoted by JV) and the versions adjusted by fpc $(1-f_1)$, $(1-f_2)$, $(1-f_3)$, $(1-f_4)$ respectively (denoted by JV1, JV2, JV3, JV4), we also used the variance estimators of \hat{Y}_{grpi} proposed by Särndal (1982) and Särndal et al. (1989).

$$VS1 = \frac{1}{2} \sum_{\substack{i, j \in S \\ i \neq j}} \frac{\Pi_i \Pi_j - \Pi_{ij}}{\Pi_{ij}} \left(\frac{\hat{e}_i}{\Pi_i} - \frac{\hat{e}_j}{\Pi_j} \right)^2 \tag{4.24}$$

$$VS2 = \frac{1}{2} \sum_{\substack{ij \in s \\ i \neq j}} \frac{\Pi_i \Pi_j - \Pi_{ij}}{\Pi_{ij}} \left(\frac{\hat{e}_i}{\Pi_i} - \frac{\hat{e}_j}{\Pi_j} \right)^2 \quad (4.25)$$

where $\hat{e}_{j'}$ is defined as:

$$\hat{e}_{j'} = \hat{e}_j - e_j \left\{ \left[\left(\hat{N} - N \sum_s x_\ell^2 / v_\ell \Pi_\ell \right) - \frac{(\hat{X} - X) \sum_s x_\ell / (\Pi_\ell v_\ell)}{v_j} \right] \right. \\ \left. \left[-(\hat{N} - N) \sum_s \frac{x_\ell^2 / (v_\ell \Pi_\ell) + (\hat{X} - X) \sum_{s'} x_\ell / (\Pi_\ell v_\ell)}{x_j / v_j} \right] \right\} \\ 1 / \left\{ \sum_s x_\ell^2 / \Pi_\ell v_\ell \sum_{\ell=1} 1 / (\Pi_\ell v_\ell) - \sum_s x_\ell / (\Pi_\ell v_\ell) \right\}^2 \quad (4.26)$$

where:

$$\hat{N} = \sum_s 1 / \Pi_\ell, \quad \hat{X} = \sum_s x_\ell / \Pi_\ell. \quad (4.27)$$

With all three sampling schemes, we selected one unit per stratum, so that estimators (4.24) and (4.25) would be zero. Hence, we used the Horvitz-Thompson form of these estimators instead:

$$VS1^* = \sum_{i \in s} \frac{1 - \Pi_i}{\Pi_i} \hat{e}_i^2 + \sum_{i, j \in s} \frac{\Pi_{ij} - \Pi_i \Pi_j}{\Pi_i \Pi_j} \hat{e}_i \hat{e}_j \quad (4.28)$$

$$VS2^* = \sum_{i \in s} \frac{1 - \Pi_i}{\Pi_i} \hat{e}_i'^2 + \sum_{i, j \in s} \frac{\Pi_{ij} - \Pi_i \Pi_j}{\Pi_i \Pi_j} \hat{e}_i' \hat{e}_j'. \quad (4.29)$$

A variance estimator of \hat{Y}_{gr} derived specifically for sampling one unit per stratum (Ouyang et al. 1991), is also included in the numerical comparison. The estimator is:

$$V0 = \frac{n}{n-1} \sum_{i=1}^n \left(\frac{\hat{e}_{i_{ji}}}{\Pi_{i_{ji}}} \right)^2 \quad (4.30)$$

where unit $\{i_{ji}\}$ is the j_i th unit drawn from the i -th stratum. The numerical study based on 2000 simulations shows (table 2) that:

- 1) VO provides the least biased estimates;
- 2) the jackknife variance estimator is usually not as good as VS1* and VS2*;
- 3) the corrected jackknife variance estimators (VJ1, VJ2, VJ3, VJ4) are usually better than VS1* and VS2* and almost as good as VO;
- 4) there is not much difference among VJ1, VJ2, VJ3, and VJ4;
- 5) VJ1 is about as good as any of the other fpc; and
- 6) VJ4 is always conservative.

This section showed that the fpc developed for \hat{Y}_{HT} can also be applied to reduce the bias of the simple jackknife variance estimator of \hat{Y}_{grpi} and those variance estimators work well.

Summary

A series of fpc were derived for sampling ppswor. If pps with replacement variance estimators must be used for ppswor sampling with \hat{Y}_{HT} , adjusting by $1-f_1$ in equation (2.13) provides the least biased variance estimators for ppswor sampling schemes selecting 1 unit/stratum. For regression estimator \hat{Y}_{pi} ,

Table 2. Mean square root of variance estimator expressed as percent of the simulation standard error.

	LOBK2			NY			RG5		
	SPSCX	SPSCXK	SPPS	SPSCX	SPSCXK	SPPS	SPSCX	SPSCXK	SPPS
VS1	95.50	94.00	94.00	76.41	88.56	101.29	93.25	93.70	96.43
VS2	95.49	94.02	93.98	76.98	89.08	102.77	93.60	93.78	96.70
VO	105.92	103.68	102.39	79.96	92.63	105.97	99.13	99.60	102.43
VJ	110.47	105.79	104.69	119.28	119.80	120.50	109.14	104.70	106.65
VJ1	101.07	96.79	95.78	114.92	115.42	116.10	103.87	99.65	101.50
VJ2	100.97	99.24	99.75	114.92	115.99	117.10	103.89	100.08	102.28
VJ3	106.24	101.75	100.68	116.37	116.88	117.56	104.97	100.70	102.59
VJ4	106.24	101.75	100.68	116.37	116.88	117.56	104.97	100.70	102.58

applying the four fpc's to a jackknife variance estimator of \hat{Y}_{grpi} results in estimators with slightly larger bias than estimator V_0 in equation (4.30) for ppswr sampling schemes selecting one unit/stratum, but they are all generally less biased than the jackknife variance estimator VJ in equation (4.19) or estimators VS1 and VS2 in equation (4.28) and (4.29). $1-f_1$ is recommended as adjustments for the jackknife variance estimator of \hat{Y}_{grpi} . The standard fpc

$$\frac{N-n}{N}$$

is desirable because it adjusts variance estimation bias substantially but is almost always conservative.

References

- Brewer, K. R. W, and Hanif, M. 1983. Sampling with unequal probabilities (Lecture Notes in Statistics. Springer-Verlag, NY 164 p.
- Cochran, W. G. 1977. Sampling techniques. 3rd ed. J. Wiley and Sons, NY 428 p.
- Kott, P. S. 1988. Model-based finite population correction for the Horvitz-Thompson estimator. *Biometrika* 75: 797-799.
- Ouyang, Z., Schreuder, H. T., and Li, J. 1991. Regression estimation under sampling with one unit per stratum. *Comm. Statist.* 20: 2431-2449.
- Särndal, C. E. 1980. A two-way classification of regression estimation strategies in probability sampling. *Can. J. Stat.* 8: 165-177.
- Särndal, C. E. 1982. Implications of survey design for generalized regression estimation of linear functions. *J. Stat. Plan. Inf.* 7: 155-170.
- Särndal, C. E., Swensson, B., and Wretman, J. H. 1989. The weighted residual techniques for estimating the variance of the general regression estimator of the finite population total. *Biometrika* 76: 527-537.
- Schreuder, H. T., Li, H. G., and Wood, G. B. 1990. Model-dependent and design-dependent sampling procedures - a simulation study. USDA For. Serv., RM For. and Range Exp. Sta. Res. Paper. RM-291. 19 pp.
- Schreuder, H. T. and Ouyang, Z. 1992. Optimal sampling strategies for weighted linear regression estimation. *Can. J. For. Res.* 22: 239-247
- Wolter, K. M. 1985. Introduction to variance estimation. Springer-Verlag, NY.
- Zahner, R., Saucier, J. R., and Meyers, R. K. 1989. Tree-ring model interprets growth declines in the southeastern United States. *Can. J. For. Res.* 19: 612-621.

The United States Department of Agriculture (USDA) prohibits discrimination in its programs on the basis of race, color, national origin, sex, religion, age, disability, political beliefs, and marital or familial status. (Not all prohibited bases apply to all programs.) Persons with disabilities who require alternative means for communication of program information (braille, large print, audiotape, etc.) should contact the USDA Office of Communications at (202) 720-2791 (voice) or (800) 855-1234 (TDD).

To file a complaint, write the Secretary of Agriculture, U.S. Department of Agriculture, Washington, DC 20250, or call (800) 245-6340 (voice) or (800) 855-1234 (TDD). USDA is an equal employment opportunity employer.



Rocky
Mountains



Southwest



Great
Plains

U.S. Department of Agriculture
Forest Service

Rocky Mountain Forest and Range Experiment Station

The Rocky Mountain Station is one of seven regional experiment stations, plus the Forest Products Laboratory and the Washington Office Staff, that make up the Forest Service research organization.

RESEARCH FOCUS

Research programs at the Rocky Mountain Station are coordinated with area universities and with other institutions. Many studies are conducted on a cooperative basis to accelerate solutions to problems involving range, water, wildlife and fish habitat, human and community development, timber, recreation, protection, and multiresource evaluation.

RESEARCH LOCATIONS

Research Work Units of the Rocky Mountain Station are operated in cooperation with universities in the following cities:

Albuquerque, New Mexico
Flagstaff, Arizona
Fort Collins, Colorado*
Laramie, Wyoming
Lincoln, Nebraska
Rapid City, South Dakota

*Station Headquarters: 240 W. Prospect Rd., Fort Collins, CO 80526